보건학 석사 학위논문

Analysis of gene-smoking interaction for $\mathbf{FEV_1}$ with repeatedly observed measurement

반복 측정 자료를 이용한 FEV_1 유전자-흡연의 상호작용 효과 분석

2016년 8월

서울대학교 보건대학원

보건학과 보건학전공

박 보 람

Analysis of gene-smoking interaction for $\mathbf{FEV_1}$ with repeatedly observed measurement

반복 측정 자료를 이용한 FEV_1 유전자-흡연의 상호작용 효과 분석

지도교수 원 성 호

이 논문을 보건학 석사 학위논문으로 제출함

2016년 5월

서울대학교 보건대학원

보건학과 보건학전공

박 보 람

박보람의 석사 학위논문을 인준함

2016년 6월

위 원 장	김 호	(인)
부 위 원 장	성 주 헌	(인)
위 원	원 성 호	(인)

Analysis of gene-smoking interaction for $\mathbf{FEV_1}$ with repeatedly observed measurement

Park Boram

Public Health Science, Department of Public Health

The Graduate School

Seoul National University

Abstract

COPD (Chronic Obstructive Pulmonary Disease) is the one of the most prevalent diseases in the old age and is characterized by chronically poor airflow. Lung function declines with age, but the rate of it with smoking, in particular, accelerates the process. It suggests that heterogeneity among individuals may be explained by difference of their genotypes and therefore interactions between genotypes and smoking are expected.

In this thesis, forced expiratory volume in 1 second (FEV1) was used to identify genetic variants related with COPD and Korea Associated Resource (KARE) cohort was selected as study population. The KARE genotype data consists with 352,228 SNPs and 8,842 samples. $\mathbf{FEV_1}$ was measured every two years and we performed genome-environment-wide interaction study (GEWIS) with linear mixed model. A

stratified analysis was conducted in accordance with the smoking experience. Age, height, BMI, time intervals and amount of smoking were included as covariates and the null hypothesis $\beta_{\mathrm{SNP}} = \beta_{\mathrm{interaction}} = 0$ were tested under the 1.64E-7 p-value.

As a result, we found 4 significant SNPs located in chromosome 17. All of these SNPs have association with *SOX9* gene which is recognized as a master regulator of cartilage formation and human skeletal dysmorphology syndrome. There are a few studies about SOX9's role on lung function and smoking. Considering that the obvious genetic factor causing gene-environment interaction is not disclosed yet, our results would support the effects of *SOX9* on FEV_1.

Keywords : linear mixed model (LMM), GEWIS (gene-environment-wide interaction study), *SOX9*, FEV1, COPD, heteroscedasticity

Student Number : 2014-23371

Contents

List of Tables

List of Figures

Introduction

COPD (Chronic Obstructive Pulmonary Disease) is the one of the most prevalent diseases in the old age and characterized by chronically poor airflow. It is expected to become the third leading cause of mortality and fifth leading cause of morbidity by the year 2020[1]. Lung function declines with age, but the rate of it with smoking, in particular, accelerates the process[2]. However pulmonary function explained by smoking is approximately 15%[3]. It suggests that heterogeneity among individuals may be explained by difference of their genotypes and therefore interactions between genotypes and smoking are expected. Previous studies have shown that genetic pathology for COPD is complex because multiple genetic factors, environment factors and gene-environment interaction exist [4-6].

It is well recognized that deficiency of α1-antitrypsin, arising from mutation of *SERPINA1* gene, predisposes to COPD[7]. However, α1-antitrypsin deficiency explains only about 1~2% of COPD patients[1]. A large number of genome wide association studies (GWAS) were conducted [1, 8, 9] to identifying the more genetic risk factor. Table 1 briefly shows COPD susceptibility loci detected by GWA study. However,

except $\alpha 1$-antitrypsin, founded significant genes were not replicated well [10, 11]. This also implies the role played by gene-environment interactions. Hancook *et al*[12] conducted genome-wide joint meta-analyses(JMA) of single nucleotide polymorphism (SNP) and SNP-by-smoking association on $\mathbf{FEV_1}$ and $\mathbf{FEV_1/FVC}$. They identified three novel genes : (1) *DNER* (2) *HLA-DQB1* and *HLA-DQA2* and (3) *KCNJ2* and *SOX9*. However, even if the results of the JMA is smaller than genome-wide significant p-value (5×10^{-8}) , the p-value of SNP-smoking interaction itself was not significant except *DNERI.* Also, Curjuric, Imboden *et al*[13] studied interaction effects between smoking and gens but the replications for the strongest SNPs in *PARK2* and *CRISP2* were not success. In spite of the importance of gene-smoking interaction effects on lung function, no genes interacting with smoking was found. Therefore, further studies are needed to detect genetic factors which interact with smoking.

In this thesis, $\mathbf{FEV_1}$ which is a diagnostic tool for COPD was used as response variable and Korea Associated Resource (KARE) cohort was used as study population. The KARE genotype data consists with 352,228 SNPs and 8,842 samples. $\mathbf{FEV_1}$ have been measured every two years from 2001 to 2005, and to find out genetic variants which affect to lung function decline, repeatedly measured $\mathbf{FEV_1}$ was analyzed

with liner mixed model. From the previous study[14], it was shown that FEV_1 distribution and decline rate are different by tobacco experience. Thus we did stratified analysis according to smoking experience.

Table 1. List of genes associated with lung function by GWAS

Symbol	Name	Chr
ADAM19[9]	A Disintegrin And Metalloproteinase Domain 19	5
AGER[9]	Advanced Glycosylation End Product-Specific Receptor	6
GSTCD[8, 15]	Glutathione S-Transferase, C-Terminal Domain Containing	4
FAM13A[9]	Family With Sequence Similarity 13, Member A	4
GPR126[9]	G protein-coupled receptor 126	6
HHIP[9, 16]	Hedgehog interacting protein	4
HTR4[9, 15]	5-Hydroxytryptamine (Serotonin) Receptor 4	5
INTS12[9]	Integrator Complex Subunit 12	4
NPNT[9]	Nephronectin	4
PID1[9]	Phosphotyrosine Interaction Domain Containing 1	2
PPT2[9]	Palmitoyl-Protein Thioesterase 2	6
PTCH1[9]	Patched 1	9
THSD4[8]	Thrombospondin, Type I, Domain Containing 4	15

Methods

1. Data

1.1 Study population

The KARE data is composed of 8842 individuals with 4117 males and 4656 females which represent general population. The survey was conducted every two years from 2001 years and the age is distributed between 40 and 69. The data obtained from period 1 to period 3 (2001 to 2005 years) was used and participants who took at least one or more spirometry measurements during follow-up time was included. As a result, among 8842 individuals 8716 individuals were included and 19953 observations were utilized for stratified analysis by smoking.

1.2 Phenotype data

Every participant in KARE cohort took a questionnaire survey, diet survey, physical examination, body composition test, X-ray examination, bone density test, electrocardiogram (EKG or ECG) and pulmonary function test (PFT) from Ansung and Ansan clinical center. The questionnaire includes demographic information (sex, age, income, and

residence), life habits (smoking, drinking, and exercise), history of disease, sentiment (depression, stress) and so on. To reduce inter-observer bias and intra-observer bias, cohort researchers were educated by standard protocol.

1.3 Genotype data

8842 study population were genotyped by Affymetrix Genome-Wide Human SNP array 5.0 with 352,228 SNPs. The genotyped data were cleaned through standard quality control (QC) procedures with following exclusion threshold : (a) Hardy-weinberg equilibrium (HWE) < 0.001 (b) Excessive missing rate > 0.05 (c) Minor allele frequency < 0.05. Next, the cleaned genotyped data was imputed with reference haplotype panels from 1000 Genomes Project[17]. SHAPEIT phasing tools was used for pre-phasing of input haplotypes and IMPUTE2 was used for main tool of imputation process[18, 19]. After QC and SNP imputation, refined KARE data has 304,245 SNPs genotyping data in 8773 samples.

2. Statistical Method

2.1 Covariate selection

We used height, BMI, sex, age, smoking status and pack year of smoking as explanatory variables. We found that smoking status and pack year of smoking are significantly related with FEV_1 and their correlation is not very high. Thus, we considered both covariates. Table 2 represents correlation between explanatory variable we used and the results were based on the first measurement.

Table 2. Correlation matrix between explanatory variables

	Height	BMI	Sex	Age	PY smoking	Smoking status
Height	1	-0.07381	-0.75424	-0.2739	0.44013	0.54436
BMI	0.07381	1	0.09691	-0.03259	-0.08365	-0.11171
Sex	-0.75424	0.09691	1	0.04998	-0.5964	-0.70588
Age	-0.2739	-0.03259	0.04998	1	0.07684	-0.05659
PY smoking	0.440136	-0.08365	-0.5964	0.07684	1	0.70845
Smoking status	0.54436	-0.11171	-0.70588	-0.05659	0.70845	1

PY smoking indicates pack year of smoking

2.2 Gene-Environment Wide Interaction Study.

Statistical analysis was performed by R programming. Stratified linear mixed model by smoking experience was used to conduct GEWIS and individuals were divided by two groups : (1) non-smokers (2) smokers; ex-smokers and current smokers. The model included time (a continuous variable that explains time difference between each $\mathbf{FEV_1}$ measurements), height, sex, BMI, age, pack year of smoking and principal component scores as fixed effects. The principal component scores was from principal component analysis to genotype data for adjusting population stratification[20]. Also, the model had random intercept, random slope about time and heterogeneous autoregressive random covariance matrix. The smokers and non-smokers had the same linear mixed model structure.

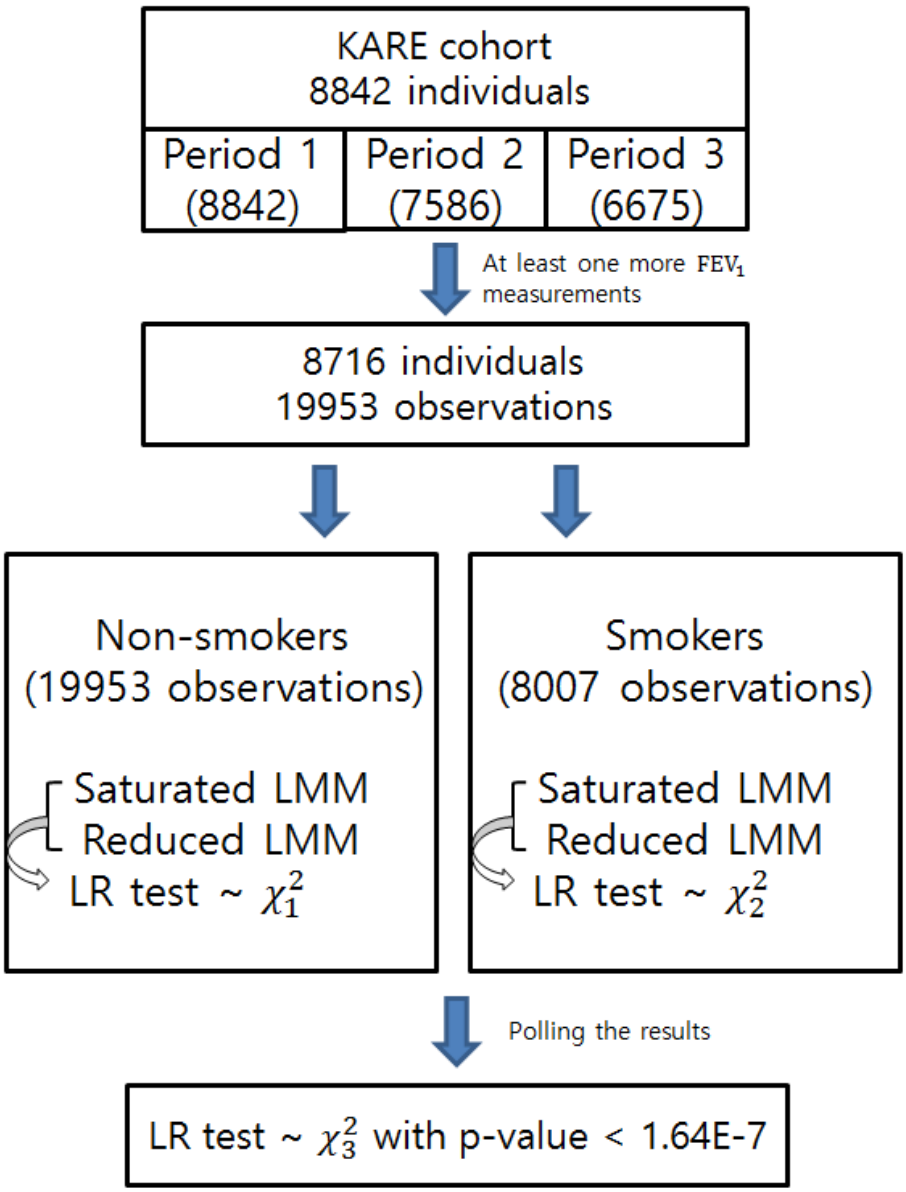
To identify specific loci which has association with $\mathbf{FEV_1}$ and which has interaction effects with smoking, log-likelihood ratio test (LRT) was utilized with the null hypothesis $\beta_{\mathbf{SNP}} = 0$ and $\beta_{interaction} = 0$. Pack year of smoking was considered as a parameter which has interaction effects with SNPs. Likelihood ratio test is widely accepted statistic for testing goodness of fit of models[21]. Likelihood ratio test statistic is $\delta = 2\log\Lambda$ where $\Lambda = \frac{\max[L_0(Under\ the\ H_0|Data)]}{\max[L_1(Under\ the\ H_1|Data)]}$ and $\mathrm{H_0}$, $\mathrm{H_1}$

represents null model (simple model) and alternative model (saturated model) respectively. Null model should be nested in alternative model. When the null hypothesis is right, the δ follows χ^2_{df} distribution (df means degree of freedom determined by difference in number of parameters between two models). We used combined likelihood ratio test to find significant SNPs for stratified linear mixed model in GEWIS. Combined likelihood ratio test can be used because stratified observations by smoking experience (1 : non-smokers , 2 : smokers) have independent relationship : If $\mathrm{X}_1, \ldots, X_k$ are independent and distributed as $\chi^2_{\mathrm{v}_1}, \ldots, \chi^2_{v_2}$, then $\mathrm{X}_1 + \cdots + X_k$ follows $\chi^2_{\mathrm{v}^*}$ where $\mathrm{v}^* = v_1 + \cdots + v_k$. Table 3 and figure 1 explains the flow of analysis.

Table 3. Analysis flow

Step 1.	The stratified linear mixed model structure.
Step 2	Likelihood ratio test (LRT) for each group. (group 1 : non-smoker, group 2 : smoker) $H_0 : \beta_{SNP} = 0$ and $\beta_{interaction} = 0$, $H_1 : not\ H_0$ Where $i = 1, \dots, 8536$ (individuals) , $j = 1,2$ (group by smoking), $k = 1,2,3$(observation numbers) Null model : $FEV1_{ijk} = \beta_0 + time_{ijk}\beta_1 + height_{ijk}\beta_2 + BMI_{ijk}\beta_3 + age_{ijk}\beta_4 + sex_{ijk}\beta_5 + sex_{ijk} \cdot age_{ijk}\beta_6 + PC1_{ijk}\beta_7 + \cdots + PC10_{ijk}\beta_{16} + b_{ij,0} + time_{ijk}b_{ij,1} + \epsilon_{ijk}$ Alternative model : $FEV1_{ijk} = \beta_0 + time_{ijk}\beta_1 + height_{ijk}\beta_2 + BMI_{ijk}\beta_3 + age_{ijk}\beta_4 + sex_{ijk}\beta_5 + sex_{ijk} \cdot age_{ijk}\beta_6 + PC1_{ijk}\beta_7 + \cdots + PC10_{ijk}\beta_{16} + SNP_i\beta_{17} + SNP_i \cdot PY_{ijk}\beta_{18} + b_{ij,0} + time_{ijk}b_{ij,1} + \epsilon_{ijk}$ ** PY means pack year of smoking For nonsmoker group, all of observation's PY value is 0 The same variance structure for two models : $\begin{pmatrix} b_{ij,0} \\ b_{ij,1} \end{pmatrix} \sim MVN(0, \begin{pmatrix} \sigma^2_{intercept} & 0 \\ 0 & \sigma^2_{time} \end{pmatrix}$ $\varepsilon_{ijk} \sim \begin{pmatrix} \varepsilon_{ij1} \\ \varepsilon_{ij2} \\ \varepsilon_{ij3} \end{pmatrix} \sim MVN(0, \Sigma_i), \Sigma_i = \begin{pmatrix} v^2_{j1} & \rho_j v_{j1} v_{j2} & \rho^2 v_{j1} v_{j3} \\ & v^2_{j2} & \rho v_{j2} v_{j3} \\ & & v^2_{j3} \end{pmatrix}$ For non-smoker group : $LRT \sim \chi^2_1$ (because no $\beta_{interaction}$ term) For smoker group : $LRT \sim \chi^2_2$
Step 3	Polled stratified LRT result. Combined LR test ~ χ^2_3

Figure 1. Analysis flow



Results

1. Characterisitics of study population

Table 4 shows basic characteristics of study population of KARE data by smoking experience. Based on the first measurement, the number of smokers and non-smokers are 3110 and 4987 respectively. Interestingly, 95% of smokers are males and 80% of non-smokers are females.

Table 4. Chracteristics of smokers

Smokers	T=1	T=2	T=3
Observed individuals	3110	3114	3844
FEV1	3.3 (±0.7)	3.3 (±0.6)	3.1 (±0.6)
Height	166.5 (±6.4)	166.8 (±6.3)	166.6 (±6.4)
BMI	24.2 (±2.9)	24.4 (±2.8)	24.3 (±2.8)
SEX			
-Male	2965 (95.3%)	2073 (95.1%)	2233 (96.0%)
-Female	145 (4.7%)	106 (4.9%)	92 (4.0%)
AGE	51.6 (±8.8)	52.4 (±8.3)	57.3 (±8.5)
Pack year of smoking	24.5 (±17.2)	23.4 (±17.9)	27.0 (±18.7)

Table 5. Chracteristics of non-smokers

Non-smokers	T=1	T=2	T=3
observed individuals	4987	2179	2325
FEV1	2.6 (±0.6)	2.6 (±0.6)	2.5 (±0.6)
Height	155.8 (±7.2)	156.6 (±7.4)	156.1 (±7.5)
BMI	24.9 (±3.2)	24.7 (±3.1)	24.7 (±3.1)
SEX			
-Male	770 (15.4%)	540 (17.3%)	714 (18.6%)
-Female	4217 (84.6%)	2574 (82.7%)	3130 (81.4%)
AGE	52.4 (±8.9)	53.2 (±8.6)	58.1 (±8.7)
Pack year of smoking	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)

2. Results of GEWIS

The FEV_1 decline rate in non-smokers was 32 mL/year for men and 27 mL/year for female. Also, in smokers, FEV_1 declie rate was 43 mL/year for men and 32mL/year for female. Smoking increases the rate of decline of FEV_1 in both males and females and p-value of smoking was less than 0.001 for men. Figure 2 represents FEV_1 decline from KARE data and Figure 3 shows distribution of FEV_1 by smoking experience. These suggest heterogeneity of FEV_1 by smoking experience.

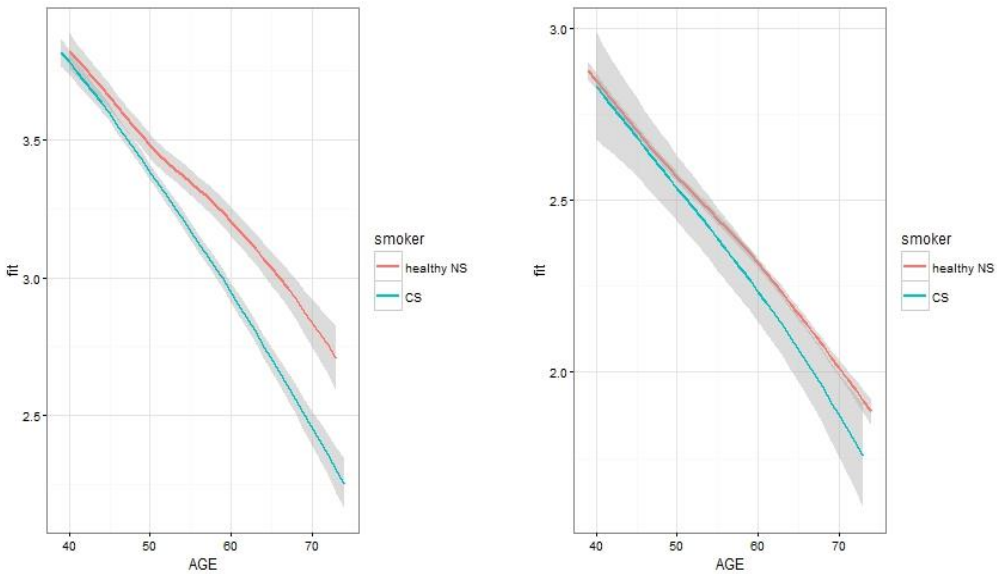
When ignoring the different trend of FEV_1 and performing GEWIS by single data set, the quantile-quantile plot was inflated [Figure 4 (a)]. However quantile-quantile plot from stratified analysis

appeared a good match between observed $-\log_{10} p$-value and those expectation under the null hypothesis [Figure 4 (b)]. As a result, we could verify that the stratified analysis follows the normality assumptions.

By using stratified linear mixed model, GEWIS analysis identified specific loci associated with $\mathbf{FEV_1}$. Four significnat SNPs were identified : rs17765644, rs17178251, rs4793541, rs11870732. The 0.05 genome-wide significance level adjusted by Bonferonni correction is 1.64E-07 and SNPs whose p-values are less than that was considered as significant. All of these four SNPs are located on chr17 and located near *SOX9* gene. *SOX9* gene has already been reported in a few previous researches[12, 22-24]. The most identified significant SNPs are shown in Table 6 with associated gene. Manhattan plot is in Figure 5 and qauntile-quantile plot is in Figure 4 (b).

According to GEWIS, *SOX9* gene reduced $\mathbf{FEV_1}$ about 55 mL for smokers and 28 mL for non-smokers. However there were no interaction effects between pack year of smoking and SNPs in smokers. The estimated SNPs effects for $\mathbf{FEV_1(\beta)}$ and p-values for each SNPs are shown in Table 7 (for non-smokers) and Table 8 (for smokers)

Figure 2. FEV_1 decline rate by smoking status

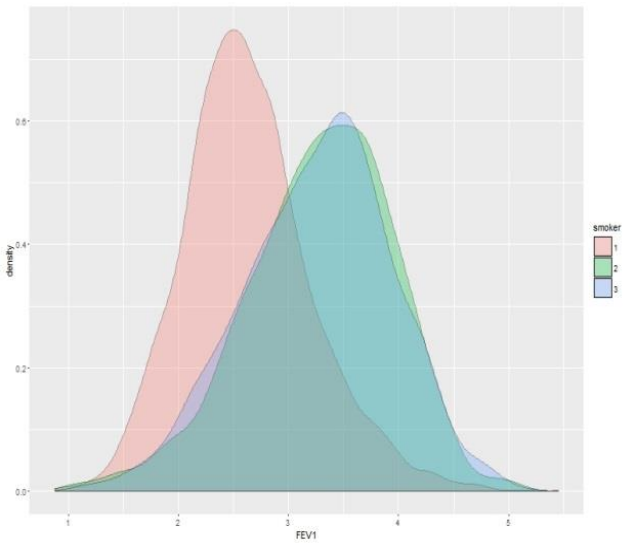


Definition of healthy NS = healthy never-smoker, CS = continuous smoker

(a) represents male's FEV_1 decline by smoking status with age.

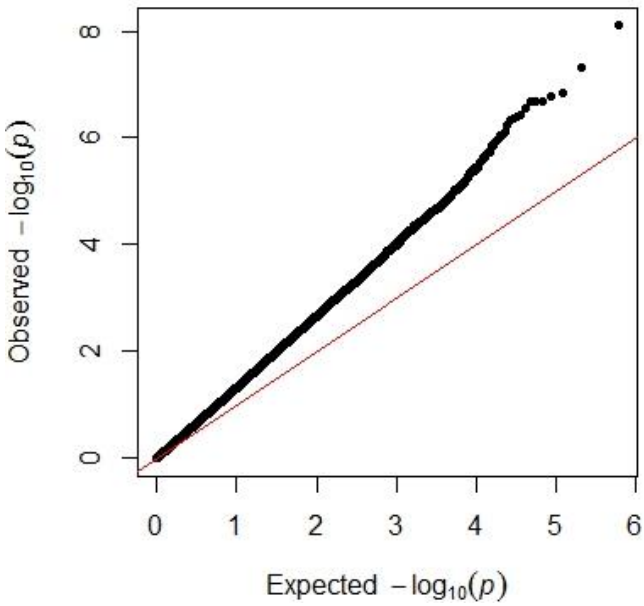
(b) represents female's FEV_1 decline by smoking status with age.

Figure 3. Distribution of FEV_1 by smoking status

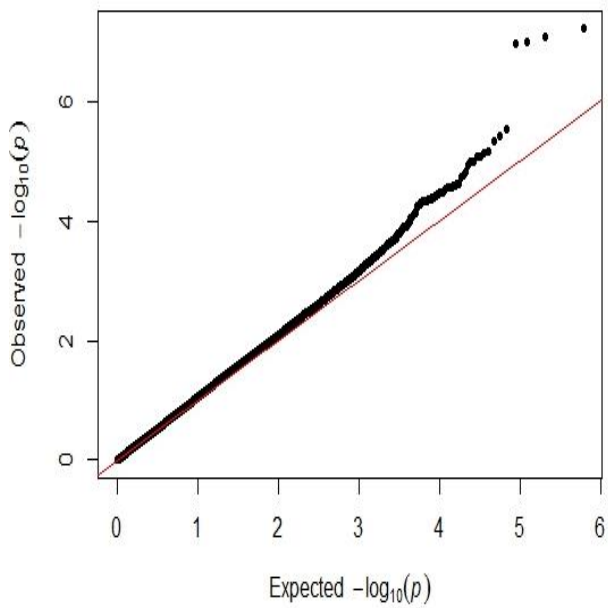


1 (red) = non-smoker(red),
2 (green) = ex-smoker
3 (blue) = current-smoker

Figure 4. Quantile-quantile plot



(a) Quantile-quantile plot from single set analysis without considering heteroscedasticity.



(b) Quanitle-quantile plot from stratififed analysis by smoking status.

Figure 5. Manhattan plot

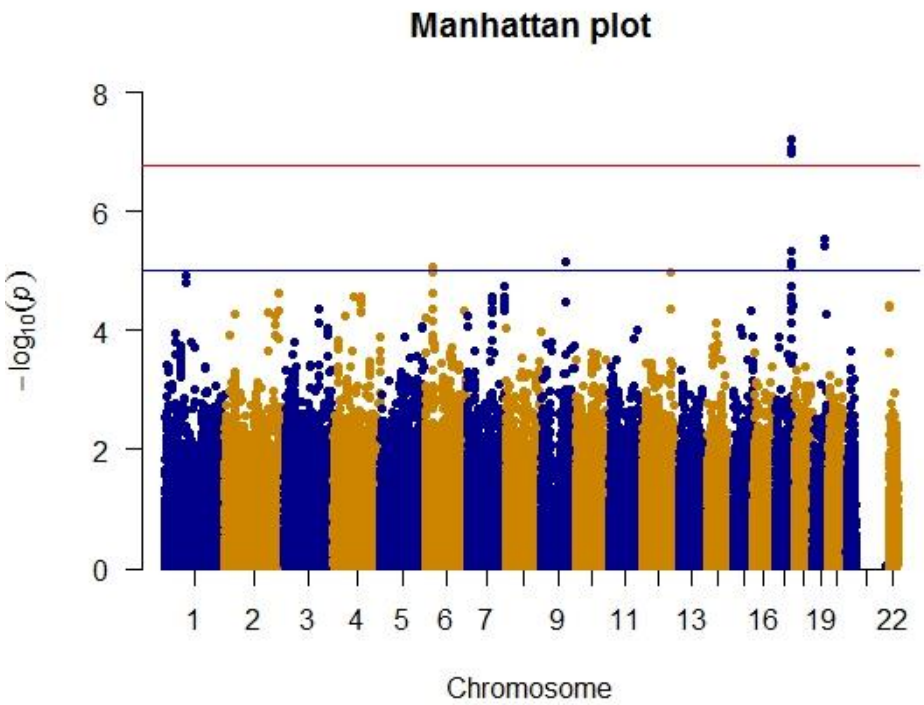


Table 6. Top 10 significant SNPs

SNP	Chromosome	Associatied Gene	Position	Minor Allele	Minor Allele Frequency	Hardy Weinberg P-value	Combined P-value
rs17765644	17	*SOX9*	69179492	C	0.3095	0.6036	**6.1E-08**
rs17178251	17		69176879	G	0.2555	0.5723	**8.4E-08**
rs11870732	17		69227595	C	0.2555	0.3214	**1.03E-07**
rs4793541	17		69195241	G	0.2555	0.6355	**1.07E-07**
rs2701	19	*CEACAM(UTR-3)*	42275099	A	0.4111	0.5223	2.86E-06
rs6508997	19		42278288	A	0.4111	0.3788	3.76E-06
rs9674957	17	*SOX9*	69111098	G	0.4204	0.2515	4.59E-06
rs700989	9	-	97515985	T	0.1222	0.31	6.84E-06
rs4328484	17	*SOX9*	69116230	G	0.4222	0.6750	7.19E-06
rs11650165	17		69109618	C	0.4166	0.5802	8.1E-06

Table 7. Estimatd effect size for most significant SNPs result for non-smokers

SNP	$\hat{\beta}$	Wald P-value	S.E
rs17765644	-0.04433	5.81E-05	0.011013
rs17178251	-0.04327	8.91E-05	0.01103
rs4793541	-0.0428	0.000101	0.010998
rs11870732	-0.04246	0.000119	0.011019
rs2701	-0.04433	5.81E-05	0.010579
rs6508997	0.016515	0.11873	0.010583
rs9674957	0.038269	0.000361	0.010719
rs4328484	0.038459	0.000366	0.010781
rs11650165	0.037805	0.000449	0.010763
rs35471	0.052776	9.64E-05	0.013519

Table 8. Estimatd effect size for most significant SNPs result for smokers

SNP	SNP Effect			SNP*Pack year of smoking Effect		
	$\hat{\beta}$	Wald P-value	S.E	$\hat{\beta}$	Wald P-value	S.E
rs17765644	-0.05679	2.60E-05	0.013485	0.000525	0.106152	0.000325
rs17178251	-0.05564	3.95E-05	0.013519	0.00052	0.110811	0.000326
rs11870732	-0.05492	4.89E-05	0.013507	0.000524	0.107744	0.000325
rs4793541	-0.05474	5.07E-05	0.01349	0.000502	0.123507	0.000326
rs2701	0.018638	0.151131	0.012981	-6.84E-05	0.825983	0.000311
rs6508997	0.018627	0.151064	0.012971	-0.0000812	0.792996	0.000309
rs9674957	0.04603	0.000417	0.013029	-0.00031	0.311145	0.000309
rs700989	0.030015	0.117931	0.019192	0.001344	0.002496	0.000444
rs4328484	0.046809	0.000356	0.013097	-0.00034	0.278019	0.000311
rs11650165	0.046429	0.000397	0.013096	-0.00035	0.260804	0.00031

Discussion

We conducted GEWIS to detect genetic risk factors significantly associated with FEV_1 and cause interaction effects with smoking. Consequently, we found four significant SNPs located on *SOX9* gene. According to the results, identified SNPs decreases FEV_1 about two times more in smokers compared to non-smokers. However it did not have interaction with pack year of smoking for smokers. These results suggests that if smokers have identified SNPs, these SNPs reduce FEV_1 regardless of amount of smoking. Further studies are needed to confirm the results.

SOX9 gene was reported from few previous studies[12, 24]. For example, Kim *et al*[24] conducted GWAS and Hancook *et al*[12] carried out SNP and SNP-by-smoking analysis through JMA. Both of them identified specific locus near *SOX9* which has associated with FEV_1. Considering that the obvious genetic factor causing gene-environment interaction is not disclosed yet, this results would support the effects of *SOX9* gene on FEV_1.

SOX9 is a master regulator of cartilage formation[25] and human skeletal dysmorphology syndrome[26]. Campomelic dysplasia is characterized by mutations in and around *SOX9*[27]. Based on the fact

that babies born with *SOX9* mutation often die in the neonatal period due to respiratory distress, studies about *SOX9*'s role on lung function had been conducted[22, 27-29]. Smoking down-regulates the canonical Wnt pathway in the airway epithelium and *SOX9* was one of the Wnt target gene[23]. These results also implies that *SOX9* has relationship with lung development and smoking.

One of the critical points in our analysis is that it showed importance of heteroscedasticity for modeling. For gene-environment wide association studies, choosing the most appropriate model is crucial[30]. In this study, $\mathbf{FEV_1}$ decline trend was significantly different by smoking status and quantile-quantile plot of homogeneous model was inflated. By classifying observations and making heterogeneity model, we could find reasonable models[31]. Even mild heteroscedasticity or mis-specified mean model could lead false contributions[32]. So the sources of heterogeneity should be identified. It is necessary to take replication study[33] with another longitudinal data to strength the importance of discovered gene and our statistical method.

However, our study has some limitations. Firstly, we may have some information bias in smoking data. The smoking information was collected by questionnaire. According to previous research[34] in 2008, there were 0.56-1.03 million hidden smoker in male and 0.8-1.65 million

hidden smoker in female in South Korea. Especially in Korea, many female tend to hide their smoking. From the KARE data, the more than 95% female responded as non-smoker. This suggests the possibility of information bias. The uncertainty of smoking information could decrease the power of stratified analysis. Secondly, the analysis did not consider other factors like existence of comorbidity, occupational dusts and socioeconomic status. One of the main reasons of difficulty in understanding COPD is that the disease come from various routes[5]. Thus, confounding factors could exist in our analysis.

In summary, according to this study, we found *SOX9* as significantly associated gene with FEV_1 and there was no interaction effect with pack year of smoking and *SOX9* gene for smokers. However we expect that by detecting significant gene on lung function, it would improve understanding of pathway to COPD and alleviate the socioeconomic burden.

Reference

1. Pillai, S.G., et al., *A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci.* PLoS Genet, 2009. **5**(3): p. e1000421.
2. Sandford, A.J., et al., *Susceptibility genes for rapid decline of lung function in the lung health study.* American journal of respiratory and critical care medicine, 2001. **163**(2): p. 469-473.
3. Silverman, E.K., et al., *Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease: risk to relatives for airflow obstruction and chronic bronchitis.* American journal of respiratory and critical care medicine, 1998. **157**(6): p. 1770-1778.
4. Walter, R., D.J. Gottlieb, and G.T. O'Connor, *Environmental and genetic risk factors and gene-environment interactions in the pathogenesis of chronic obstructive lung disease.* Environmental health perspectives, 2000. **108**(Suppl 4): p. 733.
5. Mannino, D.M. and A.S. Buist, *Global burden of COPD: risk factors, prevalence, and future trends.* The Lancet, 2007. **370**(9589): p. 765-773.
6. Sandford, A. and E. Silverman, *Chronic obstructive pulmonary disease• 1: Susceptibility factors for COPD the genotype-environment interaction.* Thorax, 2002. **57**(8): p. 736-741.
7. Stoller, J.K. and L.S. Aboussouan, *α1-antitrypsin deficiency.* The Lancet, 2005. **365**(9478): p. 2225-2236.
8. Repapi, E., et al., *Genome-wide association study identifies five loci associated with lung function.* Nature genetics, 2010. **42**(1): p. 36-44.
9. Hancock, D.B., et al., *Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function.* Nature genetics, 2010. **42**(1): p. 45-52.
10. Smolonska, J., et al., *Meta-analyses on suspected chronic obstructive pulmonary disease genes: a summary of 20 years' research.* American journal of respiratory and critical care medicine, 2009. **180**(7): p. 618-631.
11. Bossé, Y., *Updates on the COPD gene list.* International journal of chronic obstructive pulmonary disease, 2012. **7**: p. 607.
12. Hancock, D.B., et al., *Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function.* PLoS Genet, 2012. **8**(12): p. e1003098.
13. Curjuric, I., et al., *Different genes interact with particulate matter and tobacco smoke exposure in affecting lung function decline in the general population.* PloS one, 2012. **7**(7): p. e40175.
14. Wise, R.A., *The value of forced expiratory volume in 1 second decline in the assessment of chronic obstructive pulmonary disease progression.* The American journal of medicine, 2006. **119**(10): p. 4-

11.
15. Soler Artigas, M., et al., *Effect of five genetic variants associated with lung function on the risk of chronic obstructive lung disease, and their joint effects on lung function.* American journal of respiratory and critical care medicine, 2011. **184**(7): p. 786-795.
16. Zhou, X., et al., *Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP.* Human molecular genetics, 2012. **21**(6): p. 1325-1335.
17. Consortium, G.P., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-1073.
18. Delaneau, O., J. Marchini, and J.-F. Zagury, *A linear complexity phasing method for thousands of genomes.* Nature methods, 2012. **9**(2): p. 179-181.
19. Howie, B., et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.* Nature genetics, 2012. **44**(8): p. 955-959.
20. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature genetics, 2006. **38**(8): p. 904-909.
21. Posada, D. and K.A. Crandall, *Modeltest: testing the model of DNA substitution.* Bioinformatics, 1998. **14**(9): p. 817-818.
22. Rockich, B.E., et al., *Sox9 plays multiple roles in the lung epithelium during branching morphogenesis.* Proceedings of the National Academy of Sciences, 2013. **110**(47): p. E4456-E4464.
23. Wang, R., et al., *Down-regulation of the canonical Wnt β-catenin pathway in the airway epithelium of healthy smokers and smokers with COPD.* PloS one, 2011. **6**(4): p. e14793.
24. Kim, W.J., et al., *Genome-wide association studies identify locus on 6p21 influencing lung function in the Korean population.* Respirology, 2014. **19**(3): p. 360-368.
25. Arora, R., R.J. Metzger, and V.E. Papaioannou, *Multiple roles and interactions of Tbx4 and Tbx5 in development of the respiratory system.* PLoS Genet, 2012. **8**(8): p. e1002866.
26. Wright, E., et al., *The Sry-related gene Sox9 is expressed during chondrogenesis in mouse embryos.* Nature genetics, 1995. **9**(1): p. 15-20.
27. Akiyama, H., et al., *The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6.* Genes & development, 2002. **16**(21): p. 2813-2828.
28. Perl, A.-K.T., et al., *Normal lung development and function after Sox9 inactivation in the respiratory epithelium.* Genesis, 2005. **41**(1): p. 23-32.
29. Jiang, S.S., et al., *Upregulation of SOX9 in lung adenocarcinoma and its involvement in the regulation of cell growth and tumorigenicity.* Clinical Cancer Research, 2010. **16**(17): p. 4363-4373.
30. Thomas, D., *Gene-environment-wide association studies: emerging*

approaches. Nature Reviews Genetics, 2010. **11**(4): p. 259-272.
31. Verbeke, G. and E. Lesaffre, *A linear mixed-effects model with heterogeneity in the random-effects population.* Journal of the American Statistical Association, 1996. **91**(433): p. 217-221.
32. Voorman, A., et al., *Behavior of QQ-plots and genomic control in studies of gene-environment interaction.* PloS one, 2011. **6**(5): p. e19416.
33. Allen, M. and R. Preiss, *Replication and meta-analysis: A necessary connection.* Journal of Social Behavior and Personality, 1993. **8**(6): p. 9.
34. Jung-Choi, K.-H., Y.-H. Khang, and H.-J. Cho, *Hidden female smokers in Asia: a comparison of self-reported with cotinine-verified smoking prevalence rates in representative national data from an Asian population.* Tobacco Control, 2012. **21**(6): p. 536-542.

국문초록

반복 측정 자료를 이용한 FEV_1 유전자-흡연의 상호작용 효과 분석

만성 폐쇄성 폐질환(Chronic Obstructive Pulmonary Disease; COPD)는 폐에 염증이 생기는 병으로, 악화될수록 기도가 좁아지고, 폐 기능이 떨어져 숨쉬기가 어려워진다. 현재 이 질환은 2020년에는 세계 사망률 3위에 오를 것으로 예측되며, 현재 우리나라에서도 사망률 7위에 올라 있다. COPD는 대표적인 복합다인자성 질환으로 환경 요인과 유전요인에 의하여 발생한다. 폐 질환과 흡연의 관계에 대한 연구는 활발히 진행되어 왔지만, 흡연으로 인한 폐 기능의 영향은 약 15%정도만 기여하는 것으로 밝혀졌다. 또한, COPD의 대표적 유전요인인 SERPINA1은 COPD 환자의 약 1~2% 정도밖에 설명하지 못한다. 이는 흡연자 및 유전자가 COPD에 상호작용을 일으킴을 암시하지만, 현재까지 흡연과 유전자 간의 상호작용 효과에 대한 연구결과는 활발히 나오지 않고 있다.

위 논문에서는, COPD의 진단기준인 FEV_1을 종속변수로, 흡연경험, 키, BMI, 시간 및 연간 흡연량을 독립변수로 설정하였으며, 전장유전체 분석을 위해 선형혼합모형을 사용하였다. 흡연경험에 따라 FEV_1의 분포가 다름을 확인할 수 있었으며, 이에 따라 흡연에 따른 독립적 분석을 시행하였다. 이분산성을 고려하지 않은 경우에는 Q-Q이 표준 선을 따르

지 못했지만, 이분산성을 고려했을 때는 정규분포 가정을 잘 따름을 알 수 있었다. 이는 통계 모형 설정의 중요성을 시사한다고 볼 수 있다.

전장유전체 분석 결과 4개의 단일 염기 다형성(SNP)이 FEV_1에 유의한 영향을 미침을 확인할 수 있었다. 위 4개의 SNP들은 모두 염색체 17번에 위치하며, *SOX9* 유전자와 연관이 있다. 비록 흡연집단에서 SOX9과 연간 흡연량 사이의 유의한 효과는 나오지 않았지만, SOX9에 의해 FEV_1이 비흡연자 집단에서는 약 28 mL, 흡연자 집단에서는 약 55 mL 감소하는 것으로 나왔다. 추가 분석이 더 필요하지만, 이는 SOX9이 흡연유무와 관계가 있는 것으로 보인다.

SOX9 유전자는 연골 형성 및 성전환에 영향을 주는 유전자로 알려져 왔는데, 최근 *SOX9*이 폐 기능 및 흡연에 미치는 영향에 대해 연구가 진행되어 왔음을 확인할 수 있었다. 본 분석의 결과는 *SOX9*의 폐 기능 및 흡연에 대한 영향이 존재한다는 기존 연구들을 지지하고 있으며, COPD에 대한 이해도를 향상시킬 것이라 기대한다.

주요어 : 유전자-환경 전장 상호작용 분석 (GEWIS), SOX9, 만성 폐쇄성 폐질환 (COPD), , 선형혼합모형, 이분산